Journal of Computer Applications
计算机应用,2013,33(10):2846-2850,2944
ISSN 1001-9081
CODEN JYIIDU
2013-10-01
http://www.joca.cn

文章编号:1001-9081(2013)10-2846-05
doi:10.11772/j.issn.1001-9081.2013.10.2846

基于主成分分析禁忌搜索和决策树分类的异常流量检测方法

冶晓隆*,兰巨龙,郭 通
(国家数字交换系统工程技术研究中心,郑州 450002)
(*通信作者电子邮箱 yexiaolong_2854@163.com)

摘 要:真实网络流量包括大量特征属性,现有基于特征分析的异常流量检测方法无法满足高维特征分析要求。提出一种基于主成分分析和禁忌搜索(PCA-TS)的流量特征选择算法结合决策树分类的异常流量检测方法,通过PCA-TS对高维特征进行特征约减和近优特征子集选择,为决策树分类方法提供有效的低维特征属性,结合决策树分类精度和处理效率高的优点,采用半监督学习方式进行异常流量实时检测。实验表明,与传统异常检测方法相比,此方法具有更高的检测精度和更低的误检率,其检测性能受样本规模影响较小,且对未知异常可以进行有效检测。

关键词:异常检测;决策树;特征选择;主成分分析;禁忌搜索
中图分类号:TP393.08 **文献标志码**:A

Network anomaly detection method based on principle component analysis and tabu search and decision tree classification

YE Xiaolong*, LAN Julong, GUO Tong
(*National Digital Switching System Engineering and Technological R&D Center, Zhengzhou Henan* 450002, *China*)

Abstract: Real network traffic contains mass of features, and the method of anomaly detection based on feature analysis is not suitable for high-dimensional features classification. A method based on Principal Component Analysis and tabu Tabu Search (PCA-TS) decision tree classification for anomaly detection was proposed. The method reduced high-dimensional features and selected optimal feature subset which was suitable for classification through PCA-TS algorithm, then the decision tree of higher detection rate and lower false rate was used for classification and detection based on semi-supervised learning. The experiment shows that the approach has higher detection accuracy and lower false rate compared with traditional anomaly detection method, and the detection performance is less affected by sample size and is suitable for real-time detection of unknown anomalies.

Key words: anomaly detection; decision tree; feature selection; Principal Component Analysis (PCA); Tabu Search (TS)

0 引言

随着网络技术的不断发展和普遍应用,互联网安全的重要性越发凸显。网络异常中的各种攻击异常频繁发生,严重威胁着网络的正常使用。因此如何及时有效地检测网络异常,保证安全的网络环境具有重要的意义。

网络流量异常检测方法主要包括两种:统计分析[1]和机器学习[2]。基于统计的方法具有较高的检测实时性,而检测精度较低,尤其对许多隐蔽攻击无法检测;机器学习方法基于流量特征进行分析检测,由于具有较高的检测精度而成为主要研究方向。基于机器学习的异常检测主要包括聚类方法[3]和分类方法[4]:聚类方法具有无需事先样本的优点,但聚类误差导致检测精度较低;分类方法需要事先进行训练,通过训练模型进行检测,这种方法由于具有较高检测准确性而广泛使用[5-6]。基于分类的异常检测中,特征属性选择对分类精度具有重要影响[7],实际网络流量维数较高,高维数据无法应用于传统分类算法中,文献[8-10]分别采用支持向量机(Support Vector Machine, SVM)、K最近邻(K-Nearest Neighbor, KNN)和C4.5算法进行分类检测时都采用低维特征,由于其对特征属性的选择不能较好表征网络流量,造成分类精度较低,影响了检测效果。文献[8]采用SVM方法进行异常分类检测,但SVM适用于较少流量样本使得该方法无法应用于实际网络流量检测。文献[9]采用直推式的异常检测方法具有较高的检测精度,但基于"离线训练,在线检测"的机制下,由于KNN方法需要对每个样本所属类别进行判断而降低了检测效率。文献[10]利用决策树方法具有较低处理时间的特点而基于C4.5决策树算法进行异常流量实时检测,但C4.5根据信息增益率进行节点划分,由于增益值的不稳定导致分类误差较大。

基于此,本文提出了一种基于主成分分析和禁忌搜索(Principal Component Analysis and Tabu Search, PCA-TS)结合基于最短距离划分决策树(Min-Distance Decision Tree, MDDT)分类的异常流量检测方法,通过PCA-TS方法来减少高维特征空间冗余和选择最优特征子集,为分类检测提供低维和有效的流量属性,结合决策树检测实时性高的特点,该方法可以有效地进行网络流量异常实时检测。

1 相关研究

1.1 基于PCA-TS的特征选择方法

1.1.1 主成分分析算法

主成分分析(Principal Component Analysis, PCA)是统计学中分析数据的一种有效方法,主要用于特征抽取和数据降维。

收稿日期:2013-03-25;**修回日期**:2013-05-30。 **基金项目**:国家科技支撑计划项目(2012BAH02B01,2012BAH02B03);国家863计划项目(2011AA01A103,2011AA01A101,2011BAH19B04)。

作者简介:冶晓隆(1987-),男,宁夏固原人,硕士研究生,主要研究方向:网络流量异常检测; 兰巨龙(1962-),男,河北张北人,教授,博士生导师,主要研究方向:宽带信息网络; 郭通(1984-),男,江西南昌人,博士研究生,主要研究方向:网络流量测量。

其思想是利用数据集统计性质的特征空间变换,将一个数据维数较高且互相关联的数据集进行降维。通过 PCA 降维后,将原始空间转换为新的主成分空间,且各主成分互不相关。

假设含有 N 个样本的网络流量数据集 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_m\} \in \mathbf{R}^n$,其中:$\mathbf{R}^n$ 为特征空间,m 为特征维数。求得变量空间 $\boldsymbol{Z} = \{z_1, z_2, \cdots, z_k\}$,满足 $k < m$ 且 $\mathrm{cov}(z_i, z_j) = 0$,通过变换求得 k 个新变量 $\boldsymbol{Z}$ 可以代表 m 个原始变量 $\boldsymbol{X}$ 的大部分信息,即:

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{X} \tag{1}$$

其中:$\boldsymbol{\Sigma}$ 为一个 $m \times m$ 的正交矩阵,它是数据样本协方差矩阵 $\boldsymbol{C} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{u})(\boldsymbol{x}_i - \boldsymbol{u})^{\mathrm{T}}$ 的特征值矩阵,其中 $\boldsymbol{u} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$。因此转化为求解如下本征问题:

$$\lambda_i \boldsymbol{P} = \boldsymbol{C}\boldsymbol{P} \tag{2}$$

其中:λ_i 为 $\boldsymbol{C}$ 的特征值,$\boldsymbol{P}$ 为相应的特征向量。

主成分分析通过选择贡献率较大的几个特征值 λ_i 对应的特征向量 $\boldsymbol{P}$ 作为主成分,达到降维的目的。特征贡献率如下式计算:

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} = \sum_{i=1}^{m}\lambda_i / n \geqslant R \tag{3}$$

其中 R 为特征贡献率阈值,特征维数 m 选择根据 R 来确定,一般选择 R 为 85% ~ 95%。

在使用 PCA 进行分析时,由于数据中不同的变量往往有不同的量纲,会引起各变量取值的分散程度差异较大,从而影响计算精度。为了消除由于量纲的不同可能带来的影响,首先需要对变量进行标准化处理,然后利用 PCA 进行降维。

1.1.2　禁忌搜索算法

禁忌搜索(Tabu Search, TS)算法是一种启发式全局寻优搜索方法,其通过标记已搜索局部最优解和避免迭代计算中重复搜索来获得全局最优解[11]。TS 主要思想是:首先确定一个初始有效解 z,对每个解 z 定义一个邻域 $Y(z)$,从当前解的邻域中确定若干的候选解,从中选出最佳候选解。选择最佳候选解是一个搜索过程,为了避免搜索过程限于循环,TS 算法通过构造禁忌表和定义停止规则避免了搜索算法的局部最优。其中禁忌表存入前 n 次禁忌长度,避免了回到原先的解,从而提高了解空间的搜索能力;停止规则定义在若干迭代次数内最优解无法改进时,算法停止。另外禁忌搜索算法中涉及邻域、禁忌表、禁忌长度、特赦规则和初始解等都会直接影响搜索优化结果[12]。

基于禁忌搜索的特征选择是通过目标函数进行约束的最优化问题,合适的目标函数提高了搜索和最优特征选择的质量。一个好的特征解应在最少的特征数量上保证尽可能多的分类信息。在信息论理论中,一个属性的信息增益越大,其包含的信息量也越大,基于信息增益可以有效评估特征向量的分类信息,因此本文选择信息增益作为目标函数。定义目标函数如下:

$$G_T = \sum_{i=1}^{m} C(i) \times \left[\sum_{j=1}^{n} G(A_j) / n\right] \tag{4}$$

其中:$C(i)$ 为第 i 个样本是否被正确分类,m 为样本数目;$G(A_i)$ 为第 i 个特征的信息增益。通过式(4)确保特征以较少的特征数量保证最大的分类信息,选择除以 n 可以确保更快的禁忌搜索速度和避免过拟合。

禁忌搜索中初始解的选择对禁忌搜索的效果影响很大,在基于网络流量特征的最优特征选择中,由于实际网络流量特征维数较大,会影响禁忌搜索算法的效率,同时网络流量特征的冗余也对最优特征集的选择产生影响。因此禁忌搜索的初始解对搜索效率和质量具有重要影响。

1.1.3　PCA-TS 特征选择算法

特征选择是从特征集 $CT = \{c_1, c_2, \cdots, c_n\}$ 中选择一个子集 $CT' = \{c'_1, c'_2, \cdots, c'_n\}$,$c' \leqslant c$。其中:$c$ 为原始特征空间大小,c' 为特性选择后新特征空间大小。即:通过从原始特征空间中选择部分有效特征组成新的低维特征空间,其本质为一个寻优过程。

网络流量特征属性空间的“维数灾难”严重降低了基于特征分析方法的效率,而这些特征中存在大量的冗余和弱特征属性,需要通过特征约减来去除冗余和弱属性,得到精简特征属性向量。PCA-TS 方法通过 PCA 对高维特征向量进行有效降维,为禁忌搜索提供了低冗余和低维数的特征向量。结合禁忌搜索寻找近优特征子集的特点,提高了禁忌搜索的效率和精度。因此通过 PCA-TS 可以在高维特征空间中寻找最优特征子集。PCA-TS 方法具体步骤如下:

步骤 1　禁忌表置空,设置初始化参数:禁忌长度 $L_J = 13$,最大迭代次数 $D_{\max} = 600$,最大改进次数 $\bar{D}_{\max} = 100$。

步骤 2　使用 PCA 对原始网络流量特征进行约减,得到约减特征集 $T = \{T_1, T_2, \cdots, T_p\}$,$p$ 为约减后特征集数量。

步骤 3　对特征集 T 进行二进制编码,得到初始解 R_{initN}。

步骤 4　设置终止条件,当达到 $D_{\max}$ 时,搜索停止;当通过 $\bar{D}_{\max}$ 寻找最优解无改进时,停止搜索。

步骤 5　判断是否满足终止条件,如果满足终止条件,结束运算,输出最优特征子集;否则转到下一步。

步骤 6　初始解 R_{initN} 代入邻域结构计算邻域解,通过目标函数选择最佳候选解。

步骤 7　判断候选解是否满足特赦规则,如果满足,则更新禁忌表中最优解,转入步骤 4;否则转到下一步。

步骤 8　计算候选解的禁忌属性,选择非禁忌对象的最优值替换禁忌表的最初值,转入步骤 4。

步骤 9　结束,输出最优特征子集。

1.2　C4.5 决策树方法

决策树方法作为一种机器学习方法中的预测模型,代表对象属性和对象值之间的映射关系,它能从无规则的实例集合中归纳出一组采用树形结构表征的分类规则。常用的决策树方法包括:ID3 算法、CART 算法和 C4.5 算法等。与其他算法相比,C4.5 决策树方法由于具有较高的处理效率和分类稳定性,适用于网络流量的实时分类[13]而在网络流量分类中广泛使用。

决策树创建中内部节点分枝的选择是关键,对于不同划分得到的决策树的性能不同,传统 C4.5 算法利用信息熵原理,选择信息增益最大的属性作为分类属性。定义样本集 $\boldsymbol{S}$ 的理想划分 $\boldsymbol{S} = \{s_1, s_2, \cdots, s_n\}$,则信息增益率为

$$R(\boldsymbol{S}, \beta) = H_G(\boldsymbol{S}, \beta) / H_S(\boldsymbol{S}, \beta) \tag{5}$$

其中:$H_G(\boldsymbol{S}, \beta)$ 为引入测试条件 β 对 $\boldsymbol{S}$ 进行划分所得信息增益,$H_S(\boldsymbol{S}, \beta)$ 为引入条件 β 带来的分割信息熵。

C4.5 方法采用信息增益率来确定节点的分枝,文献[14]分析了采用这种方法带来的问题:划分产生的分割信息很小时,增益的值不稳定。这种不稳定可能导致信息增益率很大或者为 0,带来较大分类误差。本文采用最短距离划分方法来构

建决策树,定义 Mantaras 范氏距离[15]为两个划分间的距离,采用与理想划分距离最近的属性作为当前节点的测试条件。

定义特征属性 p_i 作为测试条件 p 得到的划分 $S' = \{s'_1, s'_2, \cdots, s'_m\}$,则理想划分 S 和划分 S' 的 Mantaras 范氏距离为:

$$d(S,S') = \frac{H(S' \mid S) + H(S \mid S')}{H(S,S')} \tag{6}$$

其中:$H(S' \mid S)$ 为理想划分 S 对划分 S' 的条件熵,$H(S \mid S')$ 为划分 S' 对理想划分 S 的条件熵,$H(S,S')$ 为理想划分 S 对划分 S' 的联合熵。

决策树训练中可能存在过度拟合,这会对新的数据集分类效果产生影响,因此要对初始决策树进行剪枝,从而得到一般的分类规则。本文利用训练数据集中剩余样本,采用悲观错误剪枝(Pessimistic Error Pruning,PEP)算法对生产初始决策树进行剪枝,PEP 算法对每棵子树只进行一次检查,具有较快的处理速度。且本方法不需要额外数据集,结合 PEP 算法可使本方法适用于样本较多数据集。

2 基于特征分类的检测模型

基于特征分类的检测模型如图 1 所示。首先对网络流量进行提取特征和数据预处理,得到待检测特征向量。离线训练阶段首先需要对高维特征空间通过特征选择进行降维,得到最优特征子集形成训练集,分类训练利用分类算法 MDDT 得到正常和异常类别,分类训练结果对检测规则库更新实现异常检测。

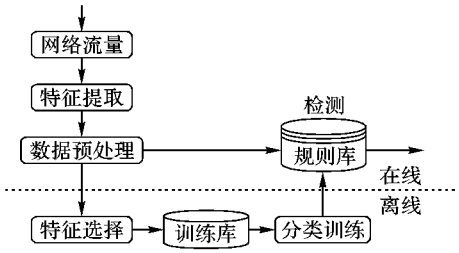

图 1 基于特征分类的检测模型

2.1 数据预处理

网络流量提取的特征中,包含不同数据类型:名词型和数值型等,且不同特征量纲也不同,这种差异会影响分类精度,所以需要将样本的属性值转换为标准的取值空间。本文对于数值型样本属性进行归一化处理,而对于如协议类型、服务类型等名词型属性根据其每个取值在取值空间的出现频次进行标准化处理。归一化方法为:

首先计算样本中每个特征属性的均值和方差:

$$mean[i] = \frac{1}{m}\sum_{k=1}^{m} sample_k[i] \tag{7}$$

$$var[i] = \frac{1}{m-1}\sum_{k=1}^{m} (sample_k[i] - mean[i])^2 \tag{8}$$

其中:$sample_k[i]$ 表示样本 k 中第 i 个属性,m 为样本数量。通过下式可得到归一化样本属性:

$$norm[i] = \frac{sample[i] - mean[i]}{\sqrt{var[i]}} \tag{9}$$

2.2 特征选择

网络流量的统计特征指的是在报文(packet)和流(flow)的属性中,抽取和端口及协议无关的特征,如报文长度、报文到达间隔时间、报文数量、流的持续时间和流中报文个数等,这些统计特征用特征矢量来表示。如一条网络流 $\boldsymbol{F}$,基于该流的特征描述可表示为 $\boldsymbol{F} = \{y_1, y_2, \cdots, y_n\}$,其中 y_i 代表特征的取值。流的特征集合可能包含多达几百个特征,通过特征选择寻找少量最优特征子集来近似描述流量对提高学习效率等具有重要意义。

在基于网络流量特征的流量分析中,一般情况下,特征数量越大,会产生更高的分析精度。但实际中,过大的特征空间会产生两个问题:1)巨大的特征空间不仅需占用更多的存储空间,而且增加了测量时间,难以应用于实时流量分析中;2)网络流量特征存在大量冗余和弱属性,这些属性不仅降低了分析精度,而且增加了算法处理的复杂度。本文采用 PCA-TS 算法,对网络流量初始特征经过 PCA 进行降维,大大减少了特征冗余和弱属性,给禁忌搜索算法提供了更优的初始解,通过禁忌搜索得到全局最优特征子集,为后续分类算法处理降低了处理时间。

基于特征选择的分类中,不同研究人员选取不同维度的特征向量,典型的选择维度包括 37[7]、36[16] 和 22[17] 等。这些特征主要包括流信息(时间、包个数、字节数),包内部时间信息,TCP/IP 控制域信息,ACK 数量,负载大小,五元组信息等。这些选取方案都是根据表征流量的常用特征如时间,长度信息进行选择,未考虑特征的贡献度及存在的冗余。

本文根据 PCA-TS 方法对高维流量特征向量进行最优特征子集选择,提取了 22 种网络流量特征作为分类训练集的特征库,与传统特征选择方法相比,去除了 TCP/IP 控制信息、ACK 信息等对网络流量表征贡献度较低的特征信息。然而在网络流量表征中,五元组信息表征存在冗余[18],而基于信息熵的源/目的 IP 地址对异常流量的表征具有较大贡献度,因此本文采用 22 个特征属性结合归一化熵的源/目的 IP 作为最终 24 个特征属性。选择的特征属性向量如图 2 所示,其中横坐标为提取的特征属性,纵坐标为 Moore 数据集中每个特征属性在数据集中所占的比例。

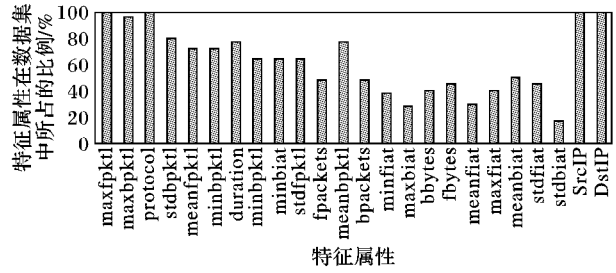

图 2 最优特征子集选择

2.3 分类训练

分类方法按照其对标记数据的依赖关系可以分为完全监督学习、无监督学习和半监督学习。完全监督学习分类准确性相对较高,但其完全依赖标记数据样本,这种方法代价昂贵无法应用于实际分类中;无监督学习一般采用聚类算法,无需标记数据进行训练,但其分类准确性较低;而半监督学习通过引入少量标记样本进行训练,不仅提高了训练器性能,而且可以对未知类型进行分类,因此本文采用半监督学习进行分类。

分类算法的选择要求具有较高分类准确性,针对网络流量大样本数据特性能有效实现分类,并且对于分类算法的实时性具有较高要求。文献[5]比较了 C4.5 和贝叶斯分类器的性能,发现 C4.5 决策树算法测试时间最短,更适合实时流量识别。本文选择基于 C4.5 的改进算法进行异常检测分类基于两点考虑:1)与 SVM 算法对于小样本的机器学习相比,C4.5 对任何样本规模都具有较好分类精度;2)C4.5 的结构可以建立方便的规则库。

利用 MDDT 算法处理分类问题通常分为两步:首先通过训练集进行学习,得到分类模型,然后通过生成的分类模型对流量进行分类。为了满足实时流量分类要求,采用"离线训练,在线识别"机制,在离线构建分类模型中,根据网络流量动态变化进行主动学习,提高分类模型的寿命和分类算法的泛化能力。

3　实验结果及分析

为了验证本文方法的有效性和可靠性,本章采用研究领域普遍使用并认可的数据集 Moore 和 KDD CUP 1999 进行实验分析。在基于特征分类的异常检测中,分类的性能对检测效果有直接影响。采用 Moore_Set 对基于 PCA-TS 的分类方法性能进行验证,通过 KDD CUP 1999 数据集对本文提出的异常检测方法性能进行分析。

3.1　实验数据和环境

3.1.1　KDD CUP 1999 数据集

为了评价本文算法对于异常检测的效果,选用 Lincoln 实验室的 KDD CUP 1999 网络数据集进行实验。该数据集包括多种网络环境下的攻击异常,主要包括 DoS、R2L、U2R 和 Probing 四类。KDD CUP 1999 数据集包括大约4 900 000条记录,4 种异常类别和正常类别(Normal)分别通过 41 个特征属性表征。

为了验证本方法的检测效果,将 KDD CUP 1999 数据集进行提取,构建三个数据集进行测试。数据集 1 包括 205 684 个正常流量数据和 2 648 个攻击异常数据;数据集 2 对数据集 1 正常数据进行提取,包括 120 000 个正常流量数据和2 648个攻击异常数据;数据集 3 对数据集 1 正常数据进行少量抽取,包括 10 000 个正常流量数据和 2 648 个攻击异常数据。三种数据集具体介绍如表 1 所示。

表 1　KDD CUP 1999 测试数据集信息

数据类别	数据集 1	数据集 2	数据集 3
DoS	1 530	1 530	1 530
Probe	900	900	900
U2R	118	118	118
R2L	100	100	100
Normal	205 684	120 000	10 000

3.1.2　实验环境及工具

本文采用的实验仿真硬件平台为普通 PC,该主机配备操作系统为 Windows XP Professional SP3,具体配置:CPU 为 Intel Core2 1.86 GHz;内存 2 GB。实验仿真软件工具采用 Matlab 2008 和 Weka-3.6.8。

本文采用异常检测方法中通用检测指标:检测率(True Positive, TP)和误报率(False Positive, FP)作为检测本方法的评价指标。其中分类算法通过准确率(precision)来表征,定义如下:

$$presicion = \frac{N_{tp}}{N_{tp} + N_{fp}} \times 100\% \qquad (10)$$

其中:N_{tp} 表示类型为 A 的网络流量样本被分类模型正确分类的数量;N_{fp} 为类型为非 A 的网络流量样本被分类模型分类为类型 A 的数量。

3.2　实验结果及分析

3.2.1　特征选择分析

对 Moore_set 数据集(样本个数为 22 932)分别取不同数量的特征进行特征分类,通过比较分类精度来分析特征数量对流量表征的影响,结果如图 3、图 4 所示。

图 3 为 Moore_Set 数据集中样本个数最多的流量类型 WWW 进行分类的精度结果。由图中可以看出,随着特征数的增加,分类精度迅速提高,当特征数增加到一定程度时,分类精度增加趋于平缓。说明网络流量可以通过较少特征属性进行表征。

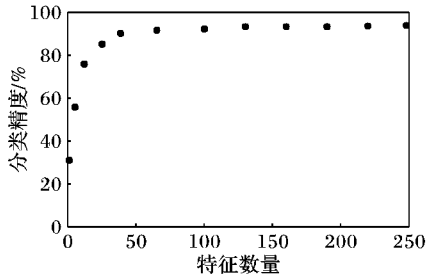

图 3　特征数量和分类精度关系

图 4 表示取不同数量特征进行分类时计算时间的比较,为了避免分类算法的影响,本实验采用 Weka 工具集中的同一分类算法 NetBayes 进行分类,其中数据样本数量为22 932。图 4 中可以看出,相同样本下分类的时间随着特征数量的增加而增加,而通过选择最优特征子集,不仅可以满足较高的分类精度,而且可以提高分类方法的效率。

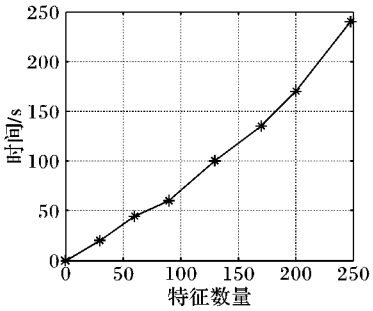

图 4　特征数量和分类时间关系

在特征选择中,常用的方法有遗传算法(Genetic Algorithm, GA)[18] 和粒子群算法[19](Particle Swarm Optimization, PSO)等。其中 TS 和 GA 方法都是利用搜索来选择最优特征子集,其相对于特征维度的处理时间如图 5 所示。从图中可以看出,随着特征维度的增加,两种方法的分析时间都成倍数地增长,但当特征维度较低时,TS 方法具有更高的处理效率。而结合 PCA 首先对网络流量高维特征进行降维,不仅去除了大量特征冗余,而且经过 PCA 降维后降低了 TS 方法的处理时间,PCA-TS 方法相对于 GA 方法针对不同特征的处理时间变化如图 6 所示。

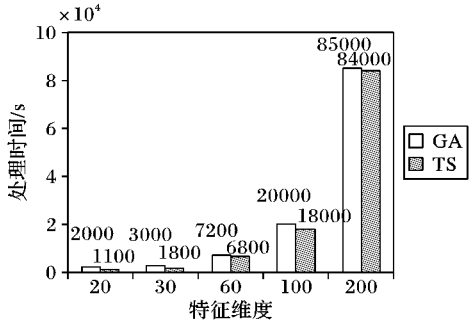

图 5　TS/GA 算法处理时间比较

从图 6 中可以看到,相比 GA 和 PSO-SVM 方法,PCA-TS 方法相对于特征维数变化时其处理时间增加较为缓慢,主要在于 PCA-TS 方法首先对高维数据进行特征约减,为禁忌搜索提供有效的低维初始特征,相比禁忌搜索,PCA 约减所耗时

间可以忽略,因此该方法具有更高的效率,可以有效地应用于网络流量的高维特征选择。

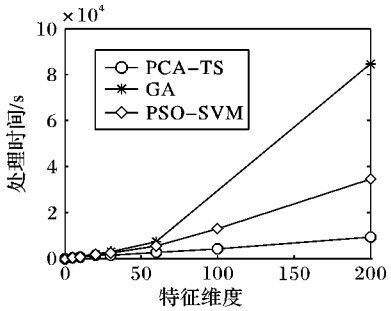

图 6 特征选择算法处理时间比较

3.2.2 检测分析

为了说明本方法的有效性,将本文方法对数据集 1 中几类攻击异常的检测效果进行分析,同时选择检测效果较好的基于直推式(TCM-KNN)的检测方法[9]、基于贝叶斯[20](NBK)的异常检测方法、基于 SVM 分类[8]检测方法及传统的 C4.5 方法[13]进行对比,结果如图 7 所示。结果表明 MDDT 方法对于四种攻击异常检测效果均优于其他四种方法,说明了本文所述方法的有效性。而对于 U2R 和 R2L 攻击异常,五种方法检测率都相对较低,主要原因在于 U2R 和 R2L 攻击与正常行为具有极大相似性,难以有效区分,但本文方法对 U2R 和 R2L 攻击异常检测效果明显提升。

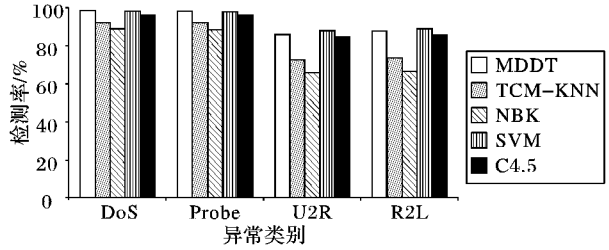

图 7 各种攻击异常检测率比较

表 2 给出了 5 种算法使用全部特征属性在三个数据集上的训练时间和分类检测时间。从训练时间上看,MDDT 和 C4.5 方法并不是最优算法;但在分类时间上,决策树仅根据样本特征属性在决策树上从上而下依次比较,处理较为简单,因此具有较高的处理效率。与传统 C4.5 方法相比,MDDT 构建的树规模更小,因此训练和检测时间更短。结合本文的“离线训练,在线检测”的处理流程,MDDT 方法适合于异常流量的实时检测。

表 2 五种算法分类检测效率对比

分类算法	训练时间/s			分类检测时间/s		
	训练集 1	训练集 2	训练集 3	训练集 1	训练集 2	训练集 3
MDDT	25.58	19.79	4.36	9.03	5.46	3.21
SVM	379.80	280.20	30.60	66.96	45.74	11.00
NBK	14.80	10.08	2.65	12.32	8.26	3.56
C4.5	30.31	21.05	5.64	10.63	6.07	3.39
KNN	380.60	279.60	28.30	229.40	102.20	15.60

在利用分类检测过程中,训练样本的数量和选取特征维数对分类器的分类效果具有一定影响。为了分析本文所述方法的鲁棒性,分别以不同训练样本规模和不同特征维数对本方法进行测试。由于数据集 2 和 3 是对数据集 1 的抽取,所以分别采用两种数据集对分类器进行训练,训练结果对数据集 1 进行检测,数据集 2 和数据集 3 具有不同的样本数量,可以充分说明训练样本在“大样本”和“小样本”情况下的检测效果。结果如表 3 所示。

表 3 训练样本规模对检测算法影响对比

检测方法	数据集 2/%		数据集 3/%	
	TP	FP	TP	FP
PCA-TS MDDT	92.63	4.63	92.51	4.55
TCM-KNN	90.33	7.51	84.73	5.73
SVM	91.31	7.93	85.34	5.42
Neural network	85.80	6.69	79.57	5.21

由表 3 可以看出,PCA-TS MDDT 方法相比其他方法具有更高的检测率和更低的误检率,并且对于训练样本变化较大时,其检测性能无明显改变。在分析特征维数对检测效果影响时,对 KDD CUP 数据集中 41 维属性特征采用 PCA-TS 方法进行特征约减,得到 14 维近优特征子集,分类算法采用两种维数特征进行训练检测,结果如表 4 所示。

表 4 特征属性数量对检测算法影响对比

检测方法	原始特征/%		特征子集/%	
	TP	FP	TP	FP
PCA-TS MDDT	92.63	4.63	92.41	4.13
TCM-KNN	90.33	7.51	89.73	7.63
SVM	91.31	7.93	91.12	8.06
Neural network	85.80	6.69	81.63	7.34

由表 4 分析可知,与其他特征检测方法相比,特征维数对本文所述方法的检测效果影响较小,然而高维特征提高了分类算法的计算时间复杂度。对于实际网络流量,特征数据维数较大,这种“维数灾难”使得分类检测变得困难,尤其是广泛使用的分类算法无法适用于高维特征和“大样本”流量异常检测。PCA-TS MDDT 方法首先对高维特征通过选择近优特征子集进行降维,再充分利用 MDDT 算法的低处理时间和高分类精度的优势。通过理论和实验分析,本文所述方法具有更高的检测精度和更低的处理时间,可以应用于网络流量异常实时检测。

4 结语

真实网络流量特征属性空间维数过高,这种“维数灾难”给基于特征分析的异常检测方法带来困难。本文提出了一种基于 PCA-TS 特征约减和决策树检测的异常流量检测方法,对网络流量特征数据通过 PCA-TS 进行特征约减降维,选择的最优特征子集作为决策树分类训练和检测的特征向量,并充分利用决策树检测实时性好的特点,采用“离线训练,在线检测”的机制,实现了异常检测的高准确率和实时性。通过 KDD CUP 1999 进行实验验证,实验表明该方法具有较高的检测率和较低的误报率,可以有效地应用于异常检测。本方法可以对未知异常进行检测,但是无法判别异常类型,结合异常检测和定位判别是下一步研究的方向。

参考文献:

[1] 李晓光, 宋宝燕, 张昕. 基于滑动多窗口的时间序列流趋势变化检测[J]. 电子学报, 2010, 38(2): 321 - 326.

[2] 朱应武, 杨家海, 张金祥. 基于流量信息结构的异常检测[J]. 软件学报, 2010, 21(10): 2573 - 2583.

[3] 肖立中, 邵志清, 马汉华, 等. 网络入侵检测中的自动决定聚类数算法[J]. 软件学报, 2008, 19(8): 2140 - 2148.

[4] 李昆仑, 黄厚宽, 田盛丰, 等. 模糊多类支持向量机及其在入侵检测中的应用[J]. 计算机学报, 2005, 28(2): 274 - 280.

(下转第 2944 页)

参考文献：

[1] HUANG H, LI B H. Automatic context induction for tone model integration in Mandarin speech recognition [J]. Journal of China Universities of Posts and Telecommunications, 2012, 19(1): 94 - 100.

[2] 黄浩,朱杰. 汉语语音识别中基于区分性权重训练的声调集成方法[J]. 声学学报,2008,33(1):1 - 8.

[3] NI C J, LIU W J, XU B. Using prosody to improve Mandarin automatic speech recognition[C]// Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari: ISCA, 2010: 2690 - 2693.

[4] TIAN Y, JIA J, WANG Y X, *et al*. A real-time tone enhancement method for continuous Mandarin speechs [C]// Proceedings of the 8th International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE, 2012: 405 - 408.

[5] LEI X, OSTENDORF M. Word level tone modeling for Mandarin speech recognition [C]// Proceedings of the 32th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2007: 665 - 668.

[6] YANG W J, LEE J C, CHANG Y C, *et al*. Hidden Markov model for Mandarin lexical tone recognition[J]. IEEE Transactions on Acoustic Speech and Signal Processing, 1988, 36(7): 988 - 992.

[7] THUBTHONG N, KIJSIRIKUL B. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2001, 9(6): 815 - 825.

[8] 曹阳,黄泰翼,徐波. 基于统计方法的汉语连续语音中声调模式的研究[J]. 自动化学报, 2004, 30(2): 191 - 198.

[9] PENG G, WANG W S. Tone recognition of continuous Cantonese speech based on support vector machines [J]. Speech Communication, 2005, 45(1): 49 - 62.

[10] WANG X H, YU Y S, WU X H. Maximum entropy based tone modeling for Mandarin speech recognition[C]// Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2010: 4850 - 4853.

[11] WEI H X, WANG X H, WU H. Exploiting prosodic and lexical features for tone modeling in a conditional random field framework [C]// Proceedings of the 33th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2008: 4549 - 4552.

[12] QIAN Y, LEE T. Tone recognition in continuous Cantonese speech using supratone models[J]. Journal of the Acoustical Society of America, 2007, 121(5): 2936 - 2945.

[13] TIAN Y, ZHOU J L, CHU M. Tone recognition with fractionized models and outlined features[C]// Proceedings of the 29th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2004: 105 - 108.

[14] SINISCALCHI S M, YU D, DENG L, *et al*. Exploiting deep neural networks for detection-based speech recognition[J]. Neurocomputing, 2013, 106(15): 148 - 157.

[15] SINISCALCHI S M, SVENDSEN T, LEE C H. A bottom-up modular search approach to large vocabulary continuous speech recognition [J]. IEEE Transactions on Audio, Speech and Language, 2013, 21(4): 786 - 797.

[16] SINISCALCHI S M, LEE C H. A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition[J]. Speech Communication, 2009, 51(11): 1139 - 1153.

[17] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1 - 27.

[18] YOUNG S, EVERMANN G, GALES M, *et al*. Hidden Markov model toolkit [EB/OL]. [2012-12-22]. http://htk.eng.cam.ac.uk/docs/docs.shtml.

(上接第2850页)

[5] WILLIAMS N, ZANDER S, ARMITAGE G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(5): 5 - 15.

[6] SOYSAL M, SCHMIDT E G. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison[J]. Performance Evaluation, 2010, 67(6): 451 - 467.

[7] LIM Y, KIM H, JEONG J, *et al*. Internet traffic classification demystified: on the sources of the discriminative power[EB/OL]. [2013-01-20]. http://security.riit.tsinghua.edu.cn/mediawiki/images/7/71/Internet_Traffic_Classification_Demystified_On_the_Sources_of_the_Discriminative_Power_slides.pdf.

[8] 徐琴珍,杨绿溪. 一种基于有监督局部决策分层支持向量机的异常检测方法[J]. 电子与信息学报,2010, 32(10): 2383 - 2387.

[9] 李洋, 方滨兴, 郭莉, 等. 基于直推式方法的网络异常检测方法[J]. 软件学报, 2007, 18(10): 2595 - 2604.

[10] PAN Z S, CHEN S C, HU G B, *et al*. Hybrid neural network and C4.5 for misuse detection [C]// Proceedings of the 2003 International Conference on Machine Learning and Cybernetics. Piscataway: IEEE, 2003, 4: 2463 - 2467.

[11] WANG Y, LI L, NI J, *et al*. Feature selection using tabu search with long-term memories and probabilistic neural networks [J]. Pattern Recognition Letters, 2009, 30(7): 661 - 670.

[12] MARINAKE M, MARINAKIS Y, DOUMPOS M, *et al*. A comparison of several nearest neighbor classifier metrics using tabu search algorithm for the feature selection problem [J]. Optimization Letters, 2008, 2(3): 299 - 308.

[13] 徐鹏, 林森. 基于C4.5决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692 - 2704.

[14] 杨哲, 李领治, 纪其进, 等. 基于最短划分距离的网络流量决策树分类方法[J]. 通信学报, 2012, 33(3): 90 - 102.

[15] LOPEZ R, MANTARAS D. A distance-based attribute selection measure for decision tree induction[J]. Machine Learning, 1991, 6(1): 81 - 92.

[16] AULD T, MOORE A W, GULL S F. Bayesian neural networks for Internet traffic classification [J]. IEEE Transactions on Neural Networks, 2007, 18(1): 223 - 239.

[17] 颜若愚, 郑庆华. 使用交叉熵检测和分类网络异常流量[J]. 西安交通大学学报, 2010, 44(6): 10 - 15.

[18] LI Y M, ZHANG S J, ZENG X P. Research of multi-population Agent genetic algorithm for feature selection[J]. Expert Systems with Applications, 2009, 36(7): 11570 - 11581.

[19] HUANG C L, DUN J F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization[J]. Applied Soft Computing Journal, 2008, 8(4): 1381 - 1391.

[20] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]// Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM, 2005: 50 - 60.